DOCUMENT RESUME

ED 402 682 EA 028 148

AUTHOR	Greene, Jay P.; Peterson, Paul E.
TITLE	Methodological Issues in Evaluation Research: The Milwaukee School Choice Plan. Occasional Paper 96-4.
INSTITUTION	Harvard Univ., Cambridge, MA. Dept. of Government.
PUB DATE	29 Aug 96
NOTE	17p.
PUB TYPE	Viewpoints (Opinion/Position Papers, Essays, etc.) (120)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Academic Achievement; *Data Analysis; Elementary Secondary Education; Participant Characteristics; Performance; Program Effectiveness; *Research Design; *Research Methodology; Research Problems; Sample Size; Sampling; *School Choice; Scientific Methodology; Scores; *Statistical Analysis
IDENTIFIERS	Milwaukee Public Schools WI; *Wisconsin (Milwaukee)

ABSTRACT

In August 1996 Jay P. Greene, Paul E. Peterson, and Jiangtao Du, with Leesa Boeger and Curtis L. Frazier, issued a report called "The Effectiveness of School Choice in Milwaukee." The report, referred to as GPDBF, presented data that indicated that low-income minority students in their third and fourth years of participation in the Milwaukee choice program performed better on standardized math and reading tests than did students who were not selected for the program. The GPDBF report explained why its results differed from those reported by a previous research team headed by Dr. John Witte. The Witte report found no effect of enrollment in choice schools on test performance. Witte, in the paper "Reply to Greene, Peterson, and Du," replied to the GPDBF report in late August 1996. This paper, a response to Dr. Witte, discusses methodological issues that affect the evaluation of school choice in Milwaukee. The paper argues that the Witte response failed to cast doubt on the GPDBF findings, and that Witte failed to justify his own analysis against reasonable criticism. The paper defends the GPDBF findings against three criticisms made by Witte: (1) that GPDBF used a mode of analysis inappropriate for educational research; (2) that GPDBF sample sizes were too small to allow for reasonable statistical inference; and (3) that missing cases biased the GPDBF results. Seven tables are included. (LMI)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 402 682

METHODOLOGICAL ISSUES IN EVALUATION RESEARCH: THE MILWAUKEE SCHOOL CHOICE PLAN

By

Jay P. Greene

and

Paul E. Peterson

Program in Education Policy and Governance

HARVARD UNIVERSITY

Occasional Paper 96-4/August 1996

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. Peterson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

BEST COPY AVAILABLE

EA 028 148

METHODOLOGICAL ISSUES IN EVALUATION RESEARCH: THE MILWAUKEE SCHOOL CHOICE PLAN

by

Jay P. Greene

and

Paul E. Peterson

August 29, 1996

Paper prepared for the Program in Education Policy and Governance, Department of Government and Kennedy School of Government, Harvard University.

METHODOLOGICAL ISSUES IN EVALUATION RESEARCH:

THE MILWAUKEE SCHOOL CHOICE PLAN

In mid-August, 1996 Jay P. Greene, Paul E. Peterson and Jiangtao Du, with Leesa Boeger and Curtis L. Frazier, issued a report on "The Effectiveness of School Choice in Milwaukee." This paper, hereinafter referred to as GPDBF, reports results from an analysis of data from a randomized experiment indicating that low-income, minority students, in their third and fourth years, performed better on standardized math and reading tests than did students who were not selected into the program. GPDBF explain why its results differ from those reported by an earlier research team headed by John Witte, which purported to find no effect of enrollment in choice schools on test performance. On August 26, 1996, John Witte issued a paper, "Reply to Greene, Peterson and Du," which responded to our study with heated rhetoric, incorrect facts, and unsupported reasoning.

In this paper we have chosen to discuss methodological issues that bear directly on the evaluation of school choice in Milwaukee. We shall show that nothing in the Witte response casts doubt on the findings reported in the GPDBF paper.

Witte's response makes little effort to defend his own analysis of the Milwaukee choice experiment against the numerous criticisms raised by GPDBF. The response does not deny that the Witte research team compared low-income, minority choice students to a more advantaged cross-section of Milwaukee public school students. It does not justify the assumptions the Witte team had to make in order to estimate school effects by means of linear regression on this particular data set. It does not deny that the response rate for the data used in Witte's main regression analyses relied upon a data set that had more than 80 percent of its cases missing and in which the evidence that the missing cases contaminated the analysis is very strong. It does not deny that many of the regressions he used employ a measure of family income--student participation in the subsidized school lunch program--that other data in the evaluation reveal to be a very poor proxy for family income.

Unable to justify his own analysis against reasonable criticism, Witte offers instead three criticisms of the GPDBF research design: 1) that GPDBF use a mode of analysis inappropriate for educational research; 2) that GPDBF sample sizes were too small to allow for reasonable statistical inference; and 3) that missing cases biased the GPDBF results.

Medical Experiments and Education Experiments

Witte claims that randomly assigning subjects to treatment and control groups is "used primarily in controlled medical experiments [but] it is theoretically inappropriate for modeling educational achievement..." Why randomized experimental data is not appropriate in education research is never explained. It is true that the opportunity to analyze data from randomized experiments in education is seldom available, but it is generally agreed among both social and physical scientists that, ceteris paribus, experimental data is almost always to be preferred over non-experimental data. The Tennessee study of classroom size provides an important, recent use of data from a randomized experiment in education. It provides the most convincing evidence ever produced that students learn more in smaller classes.

Witte's criticisms of GPDBF's use of this methodology reveal a lack of knowledge about the way in which one appropriately analyzes data from a randomized experiment. Analysis of data

must be done in a way that models as closely as possible the real-world nature of the experiment. In this case, Wisconsin state law required the private schools in the experiment to accept students at random when classes were oversubscribed.

Random admission was offered not to applicants to the program as a whole but to applicants to particular schools for specific grades in a given year. There was not one grand lottery but many little lotteries. A valid statistical model needs to approximate the real-world nature of these multiple lotteries. To do this, statistical analysis must "block" the data by introducing what is known as a dummy variable for every combination of the relevant categories: nine grades, three choice schools (to which more than 80 percent of the students applied), and four years during which applications were received.

Unfortunately, the data available do not identify the particular choice school to which a student applied. But because most Hispanics applied to one school, and most African Americans applied to the other two choice schools admitting most of the students, GPDBF used ethnicity as a proxy for the school to which a student applied.

Given that there were 9 grades (K-8), two ethnic groups serving as a proxy for schools, and four years in which students could apply (1990-93) there were potentially as many as 72 lotteries in which students were assigned to treatment and control groups. Since assignment is only random within each of these 72 lotteries or "blocks," it is necessary to control for them by inserting into a regression equation as many as 72 dummy variables representing each of these blocks. In practice, not every grade, in every school, in every year was oversubscribed, so there were fewer than 72 lotteries and therefore fewer than 72 dummy variables in each regression. This procedure may be familiar to some readers as a least squares dummy variable analysis.

The logic of "blocking," or controlling with dummies for the 72 lotteries in which students were assigned to treatment or control groups, seems to have escaped Witte when he writes: "In this study they `block' on race and grade. Why? Why not gender? Why not income? Why not parent education? All these variables have been demonstrated by prior research to be related to achievement." The answer to these questions is that blocking is designed to adjust for the fact that random assignment did not occur between the entire choice and non-select populations and instead occurred within 72 possible small lotteries. Inserting these dummy variables into the regression analyses is not done because they are hypothesized to be related to achievement, but because they must be controlled to compare those randomly assigned to treatment and control groups. Controlling or blocking for any other variable is not required when analyzing random experimental data.

Or to put it another way, one blocks the data not to control for antecedent characteristics--they have been taken into account through random assignment to treatment and control groups--but to model statistically the real-world nature of the randomized experiment.

But was assignment to treatment and control groups truly at random? Witte does not raise this quite reasonable question, but others might. To see whether there is reason to doubt that schools followed the law and accepted students at random, the background characteristics of treatment and control groups were compared (See Table 1). The information on background characteristics reported in this table are consistent with the assumption that the treatment and control groups were similar in essential respects. Although modest differences in mothers' education are evident, no significant differences were observed in initial test scores, family

income, parental marital status, or AFDC dependency. The ethnic composition and grade-level to which the student had applied were blocked, taking into account observed differences.

In short, there is no reason to doubt the assumption that the treatment and control groups were similar in all respects except that some won the lottery and attended private school while others lost and returned to the Milwaukee Public Schools (MPS). Based on this assumption, GPDBF's main analysis (Table 2) provides the strongest evidence of the effects of school choice. Randomization allows us to minimize the potential bias introduced by the larger number of missing cases that result from the use of controls for background characteristics.

GPDBF nonetheless conducted additional analyses to see whether the size of the estimated effects observed in the main analysis would prove robust when prior test scores and other background characteristics were taken into account. These analyses were conducted in order to see whether there was any evidence that the experiment was less than entirely random and/or whether missing cases had biased the results.

In one analysis GPDBF controlled for family income and mother's education. The sample size upon which this analysis is based is greatly reduced, because demographic information was available for fewer than 40 percent of those surveyed. Because the case base is small, the results are not statistically significant. What is instructive about the results is their close similarity to the results reported in the main analysis, indicating that the main analysis is robust even when controlling for demographic information. In a second analysis, GPDBF reports the results when test scores prior to entering choice schools are controlled. Once again, the results are reported to see whether the findings in the main analysis were robust. Though the case base is smaller because most students have no test score from the year prior to their application to the choice program, the estimated effects of schools on test performance reported in the main analysis were, on the whole, supported.

Let us repeat: Analysis of randomized experimental data does not require controls for background characteristics or test scores. Such controls are necessary only when one doubts that the experimental data are truly random. The fact that the estimated effects remain essentially the same when these factors are controlled lends further weight to the conclusion that the results reported in the main analysis are based on a data set in which no critical departures from randomness seem to have occurred.

Witte suggests that our methods were not adequately explained. The original statement of the methods used by GPDBF is found in pages 6-9 of the report. The report also refers readers to two sources on how to analyze randomized block experimental data in footnote 15 of the report. To be fair to Professor Witte, the early draft of the report sent to him did not include this note. We apologize. The methods employed were recommended to us by Donald Rubin, well known for his analyses of experimental data. After reading GPDBF, Rubin found the analysis to be fundamentally sound. University of Chicago econometrician James Heckman, in a recent telephone conversation with Peterson, had no difficulty understanding the methodology, finding it instead to be "standard."

Sample Size

The number of cases included in the regressions reported in GPDBF's main analysis vary

between 108 and 727 cases (Table 2). Whether or not the estimates of positive effects are based upon a sufficient number of cases is determined by calculating how likely it is that positive effects of the observed magnitude would appear if the true effects were nil. As the saying among statisticians goes, the proof is in the p, the probability that a positive finding might occur simply by chance if true effects were nil.

The p values for the positive effect of enrollment in a choice school on math performance after three and four years in the program were .03 and .01, respectively. The p values for reading tests after three and four years in the program were .08 and .13, respectively.

These p values are based on the assumption that enrollment in choice schools either has no effect or positive effects. Witte objects to this assumption, saying that the p value should be estimated using a two-tailed test that assumes the effect of attending a choice school is equally likely to be positive or negative. Witte claims that GPDBF's "argument is absurd given that their coefficients go in both + and - directions." This comment displays a misunderstanding of how one chooses between one and two-tailed tests. One chooses not on results from one's own data set (which Witte has mischaracterized--GPDBF found no statistically significant negative results in the main analysis) but on the basis of evidence from prior research, which has almost never found enrollment in private schools to have a negative effect on student test scores. Studies differ only in whether they find positive or no effects. The one-tail test is thus entirely appropriate.

Witte also objects that GPDBF p values do not fall below a conventional threshold of significance, .05. The results for three and four years into the program on math tests have p values of .03 and .01, respectively, well below the .05 level. After three years the positive effect of the program on reading test scores is significant at $p < .08$, which falls within the commonly used relaxed standard of significance at the .1 level. The reading gains after four years are significant at $p < .13$. The p value gives us the odds that our results could have been produced by chance if the true effects were zero. Judging from our p values, the odds are good that choice improves test scores.

The Missing Case Problem

It is always reasonable to be concerned about missing cases, a problem in almost all social scientific research. It is entirely reasonable to wonder whether results in years three and four may be biased by the fact that not all students remain in the study into the third and fourth years. GPDBF provided information suggesting that missing cases are unlikely to have contaminated the findings (Table 3). Because Witte expresses grave concern on this question, we present here additional evidence bearing on this point.

Cases are missing from the analysis for many reasons. Students were not in school on days tests were given. Students were not tested every year. Students left choice schools to go to school elsewhere; so did Milwaukee public school students. Low-income, minority families living in large central cities are a highly mobile group. Any study of this population inevitably confronts the fact that many cases will be missing from the analysis.

Missing cases may, but do not necessarily, contaminate an analysis. If cases fall out of the analysis randomly, then no bias occurs. But if the attrition from the sample is correlated with some variable associated with the dependent variable (in this case, student test scores), then the

results may not be valid.

One way of estimating whether missing-case bias results is to see whether the background characteristics of the test and control groups remaining in the sample remain essentially the same. If the students remaining in the test and control groups differ significantly in their background characteristics, one has reason to fear contamination of the results. Fortunately, they do not. Table 3 reports that the effects of enrollment in choice schools for those remaining in the program did not differ significantly from the effects for all students.

Table 4 shows that choice and non-selected students who remain in the study after three years had very similar test scores prior to their application to the choice program. Also, they had similar family income, and the incidence of AFDC dependency remained much the same. Differences in ethnicity and grade to which students applied were blocked. Table 5 shows that choice students also continued to be similar to non-selected students after four years in the study.

One can directly test for missing-case bias among non-selected students by comparing the first and second year test scores of non-selected students remaining in the study with those for whom later scores are not available. If those whose scores are not available after two years had lower first and second-year scores than those remaining in the study, the results are likely to be contaminated by selective attrition. Table 6 provides evidence that no such contamination occurred.

But what about Witte's tables that attempt to show selective attrition? Witte's Table 2 does not compare the demographic characteristics of treatment and control groups, as we do in Tables 1, 4 and 5, which show no important differences between the two groups. Instead, it reports a comparison of non-selected students who have at least one test score with those for whom no test score data at all are available. The differences reported in Witte's Table 2 are modest and are probably due to differential parental response rates to the demographic survey.

Witte's Table 3 also fails to compare test and control groups. It is further plagued by the fact that in this analysis Witte "stacked" the data set, using as his unit of analysis student-years, not students. By stacking the data, one year's post test becomes next year's prior test. In addition, the performance of one student may be counted several times. The net effect of this stacking is that sample sizes are artificially large and standard errors are artificially reduced, producing significance where none exists. Furthermore, a "prior" test score may reflect a test taken several years after entering the choice program, while the "post" score may be taken a year after returning to a lower-performing MPS school.

Table 7 reports the results of an analysis comparable to Witte's but it does not rely upon data that has been stacked. The table shows that students who continued in the choice program and students who withdrew each year began with nearly identical test scores. The table shows that, for the most part, the students who withdrew had scores similar to those who remained. In only two comparisons were differences statistically significant. In one the students leaving the study had the higher test scores; in the other continuing students had higher test scores. In the other six cases, the two groups did not differ significantly. Contrary to Witte's contention, students who withdrew were not low achievers.

Conclusion

By failing to respond to GPDBF's criticism of his own analysis of the Milwaukee voucher program, Witte seems to concede the points the paper made. His claim that the methodology GPDBF employed is inappropriate is incorrect. His assertion that the number of cases is too small to warrant the inferences GPDBF draw is unsupported by the p values in the GPDBF's main analysis. His claim that missing cases contaminate the results is not supported by a detailed look at the available evidence.

GPDBF's report and this discussion of methodological issues constitute only one small part of a large body of research that looks at the effects of enrollment in public and private schools. Though much has been learned, more research needs to be done. It is our pleasure to be part of a continuing discussion on one of the most important policy issues of our day. We welcome responsible criticism from Professor Witte and any other person who wishes to download and analyze the data on the Milwaukee choice plan from the world wide web or wishes to participate in the debate in some other way. Professor Witte is perfectly within his rights to pronounce that he does "not envision responding to any subsequent research or writings these authors [GPDBF] produce." But we think the welfare of inner-city, minority children is to be too important not to be the subject of continuing discussion and research.

APPENDIX

A NOTE ON DATA AVAILABILITY

Professor Witte says GPDBF "lied" when the paper said data were not available before February 1996. He appends to his report various documents that purport to show data were ready and available for analysis prior to that time.

The facts are otherwise. In a response to repeated requests from George Mitchell of Milwaukee, Wisconsin, Witte first refused to make data available. Only when the matter became an issue under the Wisconsin Open Records Act did Witte provide the Wisconsin Department of Public Instruction with an unusable data set. Peterson purchased a copy of this data set from DPI for $712.00 and attempted to analyze the data. Essential information was missing. Peterson is willing to share his copy of the data with any serious scholars who wishes to make their own attempt to analyze these data.

After ascertaining that the data Mitchell had requested were unusable, Peterson then formally asked Witte and the Department of Public Instruction for a usable copy of the data set. This eventually produced an artful letter from the Department of Public Instruction which left unclear whether the data would or would not be made available in usable form. Peterson was asked to pay several thousand dollars for information likely to be unusable.

Shortly thereafter, Witte wrote a letter to a member of the Wisconsin state legislature, saying that he would make the data available to all scholars by the end of the summer of 1995. The data became available in February 1996.

We report these facts not to perpetuate a now out-dated dispute but only to respond to the extraordinary assertion made by Professor Witte that GPDBF had lied.

Table 1.[a] Differences Between Selected and Non-selected Students[b]

All Students for Which Tests Scores are Available	Selected Students	Non-Selected Students
Math Pre-test (Average)	39	40
Reading Pre-test (Average)	38	39
% Black	77	82
% Hispanic	20	13
% Male	44	52
Grade Applied	2.8	3.6
Students for which Both Test Score and Parent Survey Results are Available	**Selected Students**	**Non-Selected Students**
Average Score on Prior Math Test	40	38
Average Score on Prior Reading Test	39	38
% Black	80	82
% Hispanic	17	15
% Male	45	51
% Married	24	32
% AFDC	57	55
Mother's Education (High School Diploma = 4)	4.2	3.8
Family Income	$11,250	$11,500
Grade Applied	2.7	3.5

[a] Corresponds to Table 3 in the GPDBF report.
[b] All data were blocked by ethnicity. Gender differences were controlled in the main analysis. Gender, education and income differences were controlled in the second analysis.

Table 2[a]. The Main Analysis
Percentile Point Effect of Choice Schools on Student Performances on Standardized Tests, Controlling for Gender and Blocking Data by Ethnicity, Year of Entry and Grade Level

Effect of Choice School on Performance on . . .				
	Years in Choice School			
Mathematics Test	**First**	**Second**	**Third**	**Fourth**
Estimated Effect of Choice	-0.49	-0.87	4.98	11.59
Standard Error	(1.77)	(1.92)	(2.62)	(4.62)
P value < (1-tail test)	0.39	0.33	0.03	0.01
P value < (2-tail test)	0.78	0.65	0.06	0.01
Number of cases	727	568	310	110
	Years in Choice School			
Reading Test	**First**	**Second**	**Third**	**Fourth**
Estimated Effect of Choice	-0.13	-0.06	3.13	4.81
Standard Error	(1.55)	(1.68)	(2.21)	(4.17)
P value < (1-tail test)	0.47	0.49	0.08	0.13
P value < (2-tail test)	0.93	0.97	0.16	0.25
Number of cases	691	576	309	108

[a] Corresponds to Table 4 in the GPDBF report.

Table 3.[a] Comparison of Test Scores for First Two Years of Students Remaining in Choice Compared to All Students: Percentile Point Effect of Choice Schools on Student Performances on Standardized Tests Controlling for Gender and Blocking Data by Ethnicity, Year of Entry and Grade Level

Effect of Choice School on Performance on . . .	Students Remaining in Choice		All Students (From Table 4)	
	Years in Choice		Years in Choice	
Mathematics Test	**First**	**Second**	**First**	**Second**
Estimated Effect of Choice	0.81	1.23	-0.49	-0.87
Standard Error	(3.00)	(2.46)	(1.77)	(1.92)
P value < (1-tail test)	0.39	0.31	0.39	0.33
P value < (2-tail test)	0.79	0.62	0.78	0.65
Number of cases	357	353	727	568

	Students Remaining in Choice		All Students (From Table 4)	
	Years in Choice		Years in Choice	
Reading Test	**First**	**Second**	**First**	**Second**
Estimated Effect of Choice	1.75	1.80	-0.13	-0.06
Standard Error	(2.64)	(2.20)	(1.55)	(1.68)
P value < (1-tail test)	0.26	0.21	0.47	0.49
P value < (2-tail test)	0.51	0.42	0.93	0.97
Number of cases	349	356	691	576

[a] Corresponds to Table 7 in GPDBF report.

Table 4 – Differences Between Selected and Non-selected Students in the 3rd Year

	Selected	Non-selected
Math Pre-Test (Average)	41	42
Reading Pre-Test (Average)	42	40
% Black	78	75
% Hispanic	22	25
% Male	42	46
Grade Applied	2.3	3.0
% AFDC	55	52
Mother's Education (High School Diploma = 4)	4.2	3.8
Family Income	11,000	11,730

Table 5 -- Differences Between Selected and Non-selected Students in the 4th Year

	Selected	Non-selected
Math Pre-Test (Average)	40	42
Reading Pre-Test (Average)	43	40
% Black	88	62
% Hispanic	12	38
% Male	39	49
Grade Applied	1.7	2.7
% AFDC	59	45
Mother's Education (High School Diploma = 4)	4.1	3.6
Family Income	11,250	11,080

Table 6. Comparison of Non-Selected Students Remaining the Study with Non-Selected Students for Whom Data Were No Longer Available

Mathematics	First Year	Second Year
Students Remaining in Study	-1.56	.25
Standard Error	(4.21)	(4.67)
P value < (1-tail test)	.35	.48
P value < (2-tail test)	.71	.96
Number of Cases	212	143

Reading	First Year	Second Year
Students Remaining in Study	2.03	-1.02
Standard Error	(3.80)	(4.38)
P value < (1-tail test)	.30	.41
P value < (2-tail test)	.59	.82
Number of Cases	216	147

Table 7 -- Re-analysis of Table 3 from Witte's Reply
Differences Between Students Electing to Stay in Choice Program and Those Who Withdrew

	Continuing Choice	Withdrew	p value
First Math Score	39.3	39.2	.92
First Reading Score	38.3	37.6	.51
Final Tests[1]			
Math for 1991 Class	38.1	41.0	.33
Reading for 1991 Class	38.6	47.1	.00
Math for 1992 Class	38.2	35.6	.24
Reading for 1992 Class	38.5	33.2	.01
Math for 1993 Class	40.6	38.4	.25
Reading for 1993 Class	36.6	38.3	.38
Math for 1994 Class	40.7	39.2	.35
Reading for 1994 Class	37.7	37.6	.99

[1] This score represents the final test taken in the choice school by those students who withdrew. For the continuing choice group, it is their test in the specified year of the choice program.